

# The distribution of the overlapping function

Miguel Abadi \*      Rodrigo Lambert \*

## Abstract

We consider the set of finite sequences of length  $n$  over a finite or countable alphabet  $\mathcal{C}$ . We consider the function defined over  $\mathcal{C}^n$  which gives the size of the maximum overlap of a given sequence with a (shifted) copy of itself. We compute the exact distribution and the limiting distribution of this function when the sequence is chosen according to a product measure with marginals identically distributed. We give a point-wise upper bound for the velocity of this convergence. Our results holds for a finite or countable alphabet. The non-parametric distribution is related to the prime decomposition of positive integers. We illustrate with some examples.

**Running head** The distribution of the overlapping function

**Subject class** 60Axx, 60C05, 60-XX, 60Fxx, 41A25

**Keywords** recurrence, overlapping, rare event, short return, first return, Renyi entropy.

## 1 Introduction

Consider a positive integer  $n$ . Consider the space of all sequences of length  $n$  defined over a finite or countable alphabet  $\mathcal{C}$ . In this work we consider the function  $S_n$  defined over  $\mathcal{C}^n$  and taken values on  $\{0, \dots, n-1\}$ . For each string, this function gives the size of the maximum overlap of the string with a (shifted) copy of itself and zero if there is no overlap. See Definition 2.4.

The function  $S_n$  is related to the *first return* function  $T_n$  that gives the *minimum number of shifts* we have to apply to the sequence in order to find an overlap with a copy of itself through the formula  $S_n = n - T_n$ .

The relevance of the first return function (and consequently of the overlapping function) was put in evidence in the statistical analysis of the Poincare recurrence. To prove convergence of the number of occurrences of a string (say of length  $n$ ) as  $n$  diverges, to the Poisson distribution it is necessary that the string does not overlap itself [13]. Or at least, that the proportion that overlaps, with respect to  $n$  is small [3, 7]. If this is not the case, a compound Poisson distribution is the limiting law [12]. There are also some approximations for this limit [17, 18, 20].

---

\*Instituto de Matemática e Estatística, Universidade de São Paulo.

It also appears when we consider the time elapsed until the first occurrence of the string. This time is called the *hitting time*. It is known that the hitting time can be well approximated by an exponential law with parameter given by the measure of the string [14]. But when the string overlaps itself, the parameter must be corrected by a factor which is the probability that the string does not appear twice consecutively. And this probability is given by the overlapping properties of the string [1, 2, 5, 10].

Yet, it appears when consider the *return time* instead of the hitting time. This law can also be well approximated by an exponential law with parameter being the measure of the string [14]. However when the string overlaps itself the limiting law is a convex combination of a Dirac measure at the origin and an exponential law [2, 5, 7]. As in the case of the hitting time, the parameter must be corrected by the above given factor. The weight of the convex combination is again this parameter. Surprisingly, when taking expectation (but not any other moment [2]) this parameter cancels. This fact is hidden when looking at Kac's Lemma ([15]).

As far as we know, the first paper to notice that the measure of all strings that have large overlaps converges to zero was [9]. The authors proved the exponential decay of this measure when "large" means larger or equal than  $2n/3$ . That result holds for  $\psi$ -mixing processes with exponentially decaying function  $\psi$  and with finite alphabet. Later, the same was generalized in [1] to  $\phi$ -mixing processes. Here, "large" means larger or equal than a certain proportion  $Cn$  where  $C$  is a constant depending on the cardinal of the alphabet.

Let us denote with  $T_n(x_1^n) = n - S_n(x_1^n)$  the number of shifts needed to get the first overlap of an  $n$ -string  $x_1^n = (x_1, \dots, x_n)$  with itself. It was proved in [21] using Kolmogorov complexity function and independently in [8] using Shannon, Mc-Millan & Breiman's Theorem that for a stochastic process over a finite alphabet, and with an ergodic measure  $\mu$  with positive metric entropy satisfying the specification property [16], the ratio  $T_n/n$  verifies

$$\liminf_{n \rightarrow \infty} \frac{T_n(x_1^n)}{n} = 1 ,$$

for almost every sequence  $x = (x_1, x_2, \dots)$ .

This result has also been proved for a class of non-uniformly expanding maps of the interval [14] in the context of dynamical systems.

Even when the definition of  $S_n$  (and  $T_n$ ) are purely combinatorial, it is interesting to have in mind an equivalent definition from the dynamical point of view. Fixed an  $n$ -string  $x_1^n$ , the return time of  $x_1^n$  over all infinite sequence  $y_1^\infty$ , such that  $y_1^n = x_1^n$  (i.e., a cylinder indexed by  $x_1^n$ ), is defined explicitly as

$$\tau_{x_1^n}(y_1^\infty) = \inf\{t \geq 2 \mid y_{t+1}^{t+n} = x_1^n\} ,$$

(and infinite otherwise). Then

$$T_n(x_1^n) = \inf_{x_{n+1}^\infty} \tau_{x_1^n}(x_{n+1}^\infty) . \quad (1)$$

Namely, the first return (of the finite sequence  $x_1^n$ ) function  $T_n(x_1^n)$  is the infimum of the return time of  $x_1^n$  over all the realizations of the process  $x_1^\infty$  that have as initial condition  $x_1^n$ . Thus,  $T_n$  is called the first return of (the  $n$ -string)  $x_1^n$  in the dynamical literature.

A large deviation principle for  $T_n$  was successively proved in [4, 6, 11] for processes that verify different types of mixing conditions, including product measures, ergodic Markov chains, Gibbs measures with Holder continuous potential, etc. The limit of the deviation function is related to the Renyi entropy of the measure that generates the strings (see, for instance, [22] for definition and properties of the Renyi entropy). The existence of the Renyi entropies are also proved.

Studying celular automatas, [19] showed that for a counting measure over a finite alphabet, the proportion of strings with no overlap converges to a positive constant.

Until now, nothing was known about the distribution of  $T_n$  and the existence of its limit reminded unknown. Since the sequence of random variables  $T_n$  are not tight, we are lead to consider instead  $S_n = n - T_n$ . In this work we consider a product measure  $\mathbb{P}$  over  $\mathcal{C}^n$  with marginals identically distributed. Namely, the marginal of  $\mathbb{P}$  is a probability function over  $\mathcal{C}$ , which may be finite or countable. Thas is, the string  $x_1^n$  are generated by independent, identically distributed random variables. Each of this random variables has a probability distribution defined by a vector of parameters  $\theta = (p_\alpha)_{\alpha \in A}$  lying in the parameter space

$$\Theta = \{\theta = (p_\alpha)_{\alpha \in A} \mid p_\alpha \geq 0, \sum_{\alpha \in A} p_\alpha = 1\} \subset (0, 1)^A .$$

Our main result read as follows: We present explicit expressions for the probability mass function of  $S_n$  and also for its cumulate distribution  $\mathbb{P}(S_n \geq k)$ . Moreover we show their convergence to a non-degenerated limiting distribution.

The limiting probability mass function reads  $q_k = m_2^{2k} - b_k$  where  $m_2$  is the  $\ell_2$ -norm of the parametric vector  $\theta$ , namely  $\sqrt{\sum_{\alpha \in A} p_\alpha^2}$ , and  $b_k$  is a smaller order term. Thus, the limiting distribution has an exponentially decreasing tail. We observe that, as in the aforementioned case of the large deviation of  $S_n$ , the probability of  $S_n$  is also related to the Renyi entropy function  $R_H(\beta)$ , in this case at  $\beta = 1$ . We also present an explicit expression for the correction term  $b_k$ . It is also related to the Renyi entropies, this time at positive integers  $\beta$ . We also show that a similar result holds for the cumulated distribution of  $S_n$ . As an application, we show that for the uniform (counting) measure, the limiting measure of the non-overlapping strings ( $S_n = 0$ ) is related to the prime decomposition of the positive integers.

The dynamical definition of  $T_n$  (and therefore of  $S_n$ ) allows us to think that this random variables are defined in the common space of infinite sequences.

Therefore one may ask about other types of convergences. We finish the paper showing that  $S_n$  does not converges in probability to any limiting random variable  $S$ .

This paper is organized as follows. In Section 2 we introduce some notation and the basic definitions. In Section 3 we present our results and provide some examples. Section 4 presents some tools needed for the proofs. Section 5 presents the proofs of our theorems. Finally, Section 6 shows that the convergence in distribution of  $S_n$  can not be extended to convergence in probability.

## 2 Notation and definitions

We consider a probability product measure with identically distributed marginals over a finite or countable alphabet  $\mathcal{C}$ .

The symbols of  $\mathcal{C}$  are called letters. The set  $\mathcal{C}$  which we index by a set  $A$ . We put  $p_\alpha$ ,  $\alpha \in A$  for the probability of these letters. To avoid non-interesting cases we assume that  $0 < p_\alpha < 1$  for all  $\alpha$ . Thus, the letters are generated by independent identically distributed random variables.

A finite sequence of consecutive letters of length  $n$ , is called an  $n$ -string or a word of length  $n$  and is denoted with the letter  $w$ , or  $w_i$  or even  $w_{i,j}$ . When we need to describe specifically the letters of a finite or infinite sequence, namely  $(x_a, \dots, x_b)$  with  $x_i \in \mathcal{C}$  and  $0 \leq a \leq b \leq \infty$ , we write simply by  $x_a^b$ .

If  $w_i$  is a  $n_i$ -string,  $i = 1, 2, \dots, k$ , with  $n = \sum_{i=1}^k n_i$  we write  $w_1 w_2 \dots w_k$  for the  $n$ -string which consists in the concatenation of the  $n_i$ -strings  $w_1, w_2, \dots, w_k$ .

The object of our analysis is the following.

**Definition 2.1.** For a given string  $x_1^n \in \mathcal{C}^n$ , the period or the first return of  $x_1^n$ , denoted by  $T_n(x_1^n)$ , is defined by the first self-overlapping position of the string. That is,  $T_n : \mathcal{C}^n \rightarrow \{1, \dots, n\}$  with

$$T_n(x_1^n) = \min\{k \geq 1 | x_1^{n-k} = x_{k+1}^n\}, \quad (2)$$

and  $T_n(x_1^n) = n$  when the above set is empty.

The fact that  $T_n/n$  converges to one almost surely implies that  $T_n$  is not tight, therefore it is more convenient to consider the variables  $S_n = n - T_n \in \{0, \dots, n-1\}$ . In this case we have that  $S_n/n$  converges to zero almost surely.

**Definition 2.2.** We define  $S_n(x_1^n)$  as the maximum size of the self-overlap, among all the self-overlaps of the string  $x_1^n$ . Namely,

$$S_n(x_1^n) = n - T_n(x_1^n) .$$

To study the level sets  $\{S_n = k\}$  or even the cumulated sets  $\{S_n \leq k\}$ , with  $k \in \{0, \dots, n-1\}$  we will use as a tool the following sets.

**Definition 2.3.** *Let  $n$  be a positive integer. For every positive integer  $k < n$ ,  $B_n(k)$  denotes the set of strings  $x_1^n$  such that the first block of length  $k$  is whatever it is, but then this block is concatenated until to complete the  $n$  symbols. Namely*

$$x_1^n = (\underbrace{x_1, \dots, x_k}_1, \underbrace{x_1, \dots, x_k}_2, \dots, \underbrace{x_1, \dots, x_k}_{\lfloor n/k \rfloor}, \underbrace{x_1, \dots, x_r}_1),$$

with  $0 \leq r < k$ . If  $\lfloor n/k \rfloor = n/k$ , which implies that  $r = 0$ , the last string is the empty string.

We will also use the following definition.

**Definition 2.4.** *We set  $R_n(k)$  as the set of  $n$ -strings  $x_1^n \in \mathcal{C}^n$  such that  $x_1^n$  has an overlap of size  $k$ . Namely*

$$R_n(k) = \{x_1^n \in \mathcal{C}^n \mid x_1^k = x_{n-(k-1)}^n\}.$$

It is easy to see that the following "duality" holds

$$B_n(n-k) = R_n(k) \quad \forall k = 1, \dots, n-1. \quad (3)$$

Finally we put

$$m_q = \sum_{\alpha \in A} p_\alpha^q.$$

Observe that  $m_q^{1/q}$  is the  $\mathcal{L}_q$ -norm of the parametric vector  $\theta$ . Also we put  $\rho = \max\{p_\alpha \mid \alpha \in A\}$ , namely, the  $\mathcal{L}_\infty$  norm of  $\theta$ .

Without lose of generality, we can think that the entries of  $\theta$  are disposed in non-decreasing order, say:  $\theta = (p_1, p_2, p_3, \dots)$ , where  $\rho = p_1 \geq p_2 \geq p_3 \geq \dots$ .

### 3 Results

In our main theorem we show that the cumulate distribution and the probability mass function of  $S_n$ , for strictly positive integers (namely  $k \neq 0$ ), can be written as a geometric term plus a correction term. The parameter of the geometric term is given by  $m_2$ . We show also a similar result for the limiting cumulate and mass distribution functions. Finally, we present a velocity of convergence for the convergence.

To state precisely our result we need to introduce some quantities that will appear in the theorem as correction terms. The first two are related to the

distribution of  $S_n$ . The last two are the limits of the previous ones, and are related to the limiting distribution of  $S_n$ . Let

$$a_{k,n} = \mathbb{P} \left( \bigcup_{j=\lfloor n/2 \rfloor}^{n-1} R_n(j) \setminus \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_n(j) \right) + \sum_{i=k+1}^{\lfloor n/2 \rfloor - 1} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=k}^{i-1} R_{2i}(j) \right),$$

and

$$b_{k,n} = \mathbb{P} \left( \bigcup_{j=\lfloor n/2 \rfloor}^{n-1} R_n(j) \cap R_n(k) \setminus \bigcup_{j=k+1}^{\lfloor n/2 \rfloor - 1} R_n(j) \right) + \sum_{i=k+1}^{\lfloor n/2 \rfloor} \mathbb{P} \left( R_{2i}(i) \cap R_{2i}(k) \setminus \bigcup_{j=k+1}^{i-1} R_{2i}(j) \right).$$

Further

$$a_k = \sum_{i=k+1}^{\infty} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=k}^{i-1} R_{2i}(j) \right),$$

and

$$b_k = \sum_{i=k+1}^{\infty} \mathbb{P} \left( R_{2i}(i) \cap R_{2i}(k) \setminus \bigcup_{j=k+1}^{i-1} R_{2i}(j) \right).$$

Now we state our main result.

**Theorem 3.1.** *Let  $\mathbb{P}$  be a product measure over  $\mathcal{C}^{\mathbf{N}}$  with marginals identically distributed. Then, for all positive integer  $k$  and all  $n \geq 2k$*

- a)  $\mathbb{P}(S_n \geq k) = m_2^k + a_{k,n}$ .
- b)  $\mathbb{P}(S_n = k) = m_2^k - b_{k,n}$ .
- c)  $\lim_{n \rightarrow \infty} \mathbb{P}(S_n \geq k) = m_2^k + a_k$ .
- d)  $\lim_{n \rightarrow \infty} \mathbb{P}(S_n = k) = m_2^k - b_k$ .

Furthermore, for all  $2n \geq 4k$  one has  $\mathbb{P}(S_{2n} \geq k) = \mathbb{P}(S_{2n+1} \geq k)$  and  $\mathbb{P}(S_{2n} = k) = \mathbb{P}(S_{2n+1} = k)$ .

The next corollary establishes what is the measure and the limiting measure of the set of strings with non-overlap, or simply the set of "self-avoiding words" [19].

**Corollary 3.1.** *Under the hypothesis of Theorem 3.1 one has*

- a)  $\mathbb{P}(S_n = 0) = 1 - m_2 - a_{1,n}, \forall n \geq 2$ .
- b)  $\lim_{n \rightarrow \infty} \mathbb{P}(S_n = 0) = 1 - m_2 - \sum_{i=2}^{\infty} \sum_{w \in \{S_i=0\}} \mathbb{P}(w)^2 > (1 - p_1)(1 - m_2)$ .

Furthermore, the sequence  $(\mathbb{P}(S_{2n} = 0))_{n \in \mathbb{N}}$  is decreasing. More precisely

$$\mathbb{P}(S_{2n} = 0) = \mathbb{P}(S_{2n-2} = 0) - \sum_{w \in \{S_n = 0\}} \mathbb{P}(w)^2 .$$

**Remark 3.1.** By the last statement of Theorem 3.1,  $\mathbb{P}(S_{2n+1} = 0) = \mathbb{P}(S_{2n} = 0)$  for all  $n$ .

The next theorem provides the exponential rate of convergence of our main theorem.

**Theorem 3.2.** For every non-negative integer  $k$  and every positive integer  $n \geq 4k$  the following inequalities hold

$$\begin{aligned} a) \quad & |\mathbb{P}(S_n = k) - \lim_{n \rightarrow \infty} \mathbb{P}(S_n = k)| \leq C m_2^{n/2} \left( \frac{m_3}{m_2^{3/2}} \right)^k , \\ b) \quad & |\mathbb{P}(S_n \geq k) - \lim_{n \rightarrow \infty} \mathbb{P}(S_n \geq k)| \leq C m_2^{n/2} \left( \frac{m_3}{m_2^{3/2}} \right)^k \frac{m_2^{3/2}}{m_3 - m_2^{3/2}} , \end{aligned}$$

where  $C$  is a positive constant (that depends only on vector  $\theta$ ).

The next proposition presents bounds for  $a_{k,n}, b_{k,n}, a_k, b_k$ .

**Proposition 3.1.** Under the hypothesis of Theorem 3.1 one has

$$\begin{aligned} a) \quad & a_{k,n} \leq (m_2^{k+1} - m_2^n) / (1 - m_2) . \\ b) \quad & b_{k,n} \leq (m_4^{k+1}) / (1 - m_2) + \left( 2m_2^{n/2+1} / (m_2 - \rho^2) \right) (m_3 / m_2^{3/2})^k . \\ c) \quad & a_k \leq m_2^{k+1} / (1 - m_2) . \\ d) \quad & b_k \leq m_4^{k+1} / (1 - m_2) . \end{aligned}$$

The bounds in the proposition above do not establishes which one is the leading term between  $m_2^k$  and  $a_{k,n}$  or  $a_k$  in Theorem 3.1. The next proposition shows that actually, both situations can happen. (It is obvious that  $m_2^k \geq \max_{n \geq 2k} \{b_{k,n}, b_k\}$ .)

The next proposition shows us that the bound presented in Proposition 3.1c) is sharp. Moreover, it shows that, if  $m_2 \leq 1/2$ ,  $(m_2^k)_{k \in \mathbb{N}}$  is the leading term. If  $m_2 > 1/2$ , the sequence  $(m_2^k)_{k \in \mathbb{N}}$  starts above the sequence  $(a_k)_{k \in \mathbb{N}}$ , and then its tail becomes strictly smaller.

**Proposition 3.2.** Under the conditions of Theorem 3.1, there exists  $A(k)$  (that satisfies:  $\lim_{k \rightarrow \infty} A(k) = 0$ ) such that  $a(k) \geq m_2^{k+1} / (1 - m_2) - A(k)$ . Furthermore

$$a) \quad \text{If } m_2 \leq 1/2, \text{ then } m_k > a_k \text{ for all } k \in \mathbb{N}.$$

b) If  $m_2 > 1/2$ , then

1.  $a_1 < m_2$
2. There exists some  $k_0 > 0$ , for which  $a_k > m_2^k$ , for all  $k > k_0$ .

### 3.1 Examples

To exemplify the behavior of  $m_2, a_{k,n}, b_{k,n}, a_k, b_k$  we present some examples.

**Example 3.1** (Two letters alphabet). Consider the case where  $\mathcal{C} = \{0, 1\}$  and  $\theta = (p, 1-p) = (p_1, p_2)$ , where  $p_1 = p$ . Then,  $m_i = p_1^i + p_2^i$ , for  $i \in \mathbb{N}$ , and the inequalities given by Proposition 3.1 become

$$\begin{aligned} \text{a) } a_{k,n} &\leq \frac{(p_1^2 + p_2^2)^{k+1} + (p_1^2 + p_2^2)^n}{1 - p_1^2 - p_2^2} \\ \text{b) } b_{k,n} &\leq \frac{(p_1^4 + p_2^4)^{k+1}}{1 - p_1^2 - p_2^2} + \frac{2(p_1^2 + p_2^2)^{n+1}}{p_2^2} \left( \frac{p_1^3 + p_2^3}{(p_1^2 + p_2^2)^{3/2}} \right)^k. \\ \text{c) } a_k &\leq \frac{(p_1^2 + p_2^2)^{k+1}}{1 - p_1^2 - p_2^2}. \\ \text{d) } b_k &\leq \frac{(p_1^4 + p_2^4)^{k+1}}{1 - p_1^2 - p_2^2}. \end{aligned}$$

In c), notice that if  $p > 1/2$ , then  $m_2 > 1/2$ , and item b)2. of Proposition 3.2 ( $a_k > m_2^k$ ) holds for all  $k > k_0$ , where  $k_0 = \lceil \log \frac{(1-p_1^4-p_2^4)^2}{2(p_1^2+p_2^2)-1} \rceil / \lceil \log \frac{p_1^4+p_2^4}{p_1^2+p_2^2} \rceil$ .

**Example 3.2** (Uniform measure). In this example we consider a uniform product measure over the finite alphabet  $\mathcal{C} = \{1, \dots, s\}$ , so that  $\theta = (1/s, \dots, 1/s)$ . Then,  $m_i = s(1/s^i) = 1/s^{i-1}$ . Thus  $m_i = 1/s^{i-1}$ . The inequalities given by Proposition 3.1 become

$$\begin{aligned} a_{k,n} &\leq \frac{s^{n-k}}{s^n(s-1)}, & a_k &\leq \frac{1}{s^k(s-1)} \\ b_{k,n} &\leq \frac{1}{s-1} \left( s^{-(3k+2)} + 2s^{-\frac{n+k}{2}+1} \right), & b_k &\leq \frac{s^{-(3k+2)}}{s-1} \end{aligned}$$

By Proposition 3.2a), we have that in the uniform case,  $m_2$  is always the leading term.

The proportion of words of length  $n$  with no overlap is

$$\frac{s-1}{s} - \mathbb{P} \left( \bigcup_{j=n/2}^{n-1} R_n(j) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) \right) - \sum_{i=2}^{n/2-1} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=1}^{i-1} R_{2i}(j) \right). \quad (4)$$



Further

$$\bigcup_{j=n/2}^{n-1} R_n(j) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) = \bigcup_{j=1}^{n-1} R_n(j) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) .$$

By Lemma 4.4

$$\bigcup_{j=1}^{n-1} R_n(j) = \bigcup_{j=1}^{n/2} R_n(j) .$$

Thus the leftmost probability in (4) is

$$\mathbb{P} \left( R_n(n/2) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) \right) ,$$

that can be added to the rightmost term in (4). Thus

$$\mathbb{P}(S_n = 0) = \frac{s-1}{s} - \sum_{i=2}^{n/2} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=1}^{i-1} R_{2i}(j) \right) .$$

Similarly, the limiting proportion of words with no overlap is exactly

$$\frac{s-1}{s} - \sum_{i=2}^{\infty} \frac{\#\{S_i = 0\}}{s^{2i}} = \frac{s-1}{s} - \sum_{i=2}^{\infty} \frac{1}{s^i} \mathbb{P}(S_i = 0) . \quad (5)$$

Since  $\mathbb{P}(S_{2i+1} = 0) = \mathbb{P}(S_{2i} = 0)$ , the last expression becomes

$$\frac{s-1}{s} - 2 \frac{s+1}{s} \sum_{i=1}^{\infty} \frac{1}{s^i} \mathbb{P}(S_{2i} = 0) .$$

Moreover

$$\mathbb{P}(S_{2n} = 0) = \mathbb{P}(S_{2n-2} = 0) - \frac{1}{s^n} \mathbb{P}(S_n = 0) .$$

## 4 Tools for the proofs

Before proving our main theorem, we prove a number of preparatory lemmas. Firstly, we recall the following classical notation. For a positive integer  $x$  we write  $\lfloor x \rfloor$  for the largest integer smaller or equal than  $x$ . Similarly, we write  $\lceil x \rceil$  for the smallest integer larger or equal than  $x$ .

**Lemma 4.1.** *Let  $p \geq 1$ ,  $q \geq 1$ . Then*

$$m_{qp} \leq m_q^p .$$

*Proof.* Since  $m_q^{1/q}$  is the  $\mathcal{L}_q$  norm of the vector  $\theta$ , a classical  $\mathcal{L}_q$  inequality gives

$$m_{qp} = (m_{qp}^{1/qp})^{qp} \leq (m_q^{1/q})^{qp} = m_q^p .$$

□

The following lemma is a tool to present explicit computations for the probability  $\mathbb{P}(B_n(j))$ .

**Lemma 4.2.** *The following equality holds for every positive integers  $j$  and  $\ell$*

$$\sum_{w \in \mathcal{C}^j} \mathbb{P}(w)^\ell = m_\ell^j .$$

*Proof.* For each  $w \in \mathcal{C}^j$  one has

$$\mathbb{P}(w) = \prod_{\alpha \in A} p_\alpha^{j_\alpha} , \quad \text{where} \quad \sum_{\alpha \in A} j_\alpha = j .$$

Thus

$$\mathbb{P}(w)^\ell = \prod_{\alpha \in A} (p_\alpha^{j_\alpha})^\ell = \prod_{\alpha \in A} (p_\alpha^\ell)^{j_\alpha} .$$

Thus

$$\sum_{w \in \mathcal{C}^j} \mathbb{P}(w)^\ell = \sum_{\sum_{\alpha \in A} j_\alpha = j} \binom{j}{\prod j_\alpha} \prod_{\alpha \in A} (p_\alpha^\ell)^{j_\alpha} = \sum_{\alpha \in A} p_\alpha^\ell = m_\ell^j .$$

□

The next lemma says that, the total measure of the  $n$ -strings that have small overlap remains the same if we "cut" the central letters of the strings.

**Lemma 4.3.** *Let  $k \leq \lfloor n/2 \rfloor - 1$ . Then*

$$\mathbb{P} \left( \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_n(j) \right) = \mathbb{P} \left( \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_{2(\lfloor n/2 \rfloor - 1)}(j) \right) .$$

*Proof.*  $w = x_1^n \in \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_n(j)$  if and only if there exists a  $j$  such that  $k \leq j \leq \lfloor n/2 \rfloor - 1$  and  $x_1^n \in R_n(j)$ . Thus

$$w = w_1 w_2 w_1 ,$$

where  $w_1$  is a  $j$ -string and  $w_2$  is an  $n - 2j$ -string and they are independent. Now we write  $w_2 = w_{2,1} w_{2,2} w_{2,3}$  where  $w_{2,2}$  is the central word of  $w_2$ , of length 2 in the case that  $n$  is even or of length 3 in the case that  $w_2$  it is odd. Namely

$$w_{2,2} = x_{\lfloor n/2 \rfloor}^{\lfloor n/2 \rfloor + 1} ,$$

and  $w_{2,1}$  and  $w_{2,3}$  are words of length  $\lfloor (n - 2j)/2 \rfloor - 1$ . Now, define  $\tilde{w} = w_1 w_{2,1} w_{2,3} w_1 \in R_{2\lfloor n/2 \rfloor - 1}(j)$ , which is independent of  $w_{2,2}$ . Thus

$$\begin{aligned} \mathbb{P} \left( \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_n(j) \right) &= \sum_{w \in \bigcup_{j=k}^{\lfloor n/2 \rfloor - 1} R_n(j)} \mathbb{P}(w) \\ &= \sum_{w_1 w_{2,1} \in \mathcal{C}^{n-2}} \sum_{w_{2,2} \in \mathcal{C}^i} \mathbb{P}(w_1 w_{2,1} w_{2,1} w_1) \mathbb{P}(w_{2,2}) . \end{aligned}$$

Summing independently each term, the first term sums up to  $R_{2\lfloor n/2 \rfloor - 1}(j)$  and the second one sum up to one.  $\square$

The next lemma says that the total measure of the set of  $n$ -strings with large overlap, goes to zero exponentially fast.

**Lemma 4.4.** *The following holds*

$$\mathbb{P} \left( \bigcup_{j=1}^{\lfloor n/2 \rfloor} B_n(j) \right) = \mathbb{P} \left( \bigcup_{j=\lfloor n/2 \rfloor}^{n-1} R_n(j) \right) \leq \frac{n}{2} m_2^{\lfloor n/2 \rfloor} .$$

*Proof.* The equality follows by duality. To prove the inequality, firstly we have

$$\mathbb{P} \left( \bigcup_{j=1}^{\lfloor n/2 \rfloor} B_n(j) \right) \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \mathbb{P}(B_n(j)) . \quad (6)$$

Still, if  $w \in B_n(j)$ , then we can write  $n = j\lfloor n/j \rfloor + r$  where  $0 \leq r < j$ . Thus

$$w = \underbrace{w_j w_j \dots w_j}_{\lfloor n/j \rfloor \text{ times}} w_r ; \quad w_j \in \mathcal{C}^j, \quad w_r \in \mathcal{C}^r .$$

Therefore, by Lemma 4.2

$$\mathbb{P}(B_n(j)) \leq \sum_{w_j \in \mathcal{C}^j} \mathbb{P}(w_j)^{\lfloor n/j \rfloor} \rho^r = m_{\lfloor n/j \rfloor}^j \rho^r .$$

By Lemma 4.1

$$m_{\lfloor n/j \rfloor}^j \leq m_2^{(n-r)/2} .$$

Observe that  $\rho \leq m_2^{1/2}$ . Thus, the sum in (6) is bounded from above by

$$\sum_{j=1}^{\lfloor n/2 \rfloor} (m_2^{1/2})^n = \lfloor \frac{n}{2} \rfloor m_2^{n/2} .$$

$\square$

**Lemma 4.5.** *The following holds*

$$\bigcup_{k=1}^{n-1} R_n(k) = \bigcup_{k=1}^{k=\lceil n/2 \rceil} R_n(k).$$

*Proof.* Let  $k \geq \frac{n}{2}$ . If  $\omega \in R_n(k) = B_n(n-k)$ , so  $\omega = \omega_1 \omega_1 \cdots \omega_r$ , with  $n = \left\lfloor \frac{n}{n-k} j \right\rfloor + r$ , with  $r$  being the size of  $\omega_r$  (which could be 0), for some integer  $j$ . If  $r = 0$ , we have that  $\omega$  overlaps in (at least) a  $\omega_1$  string. If  $r > 0$ , we have that  $\omega$  overlaps in (at least) a  $\omega_r$  string. In both cases, we have a smaller overlap than  $n/2$ , and it proves that  $\bigcup_{k=1}^{n-1} R_n(k) \in \bigcup_{k=1}^{k=\lceil n/2 \rceil} R_n(k)$ , and this concludes the proof.  $\square$

## 5 Proofs

### 5.1 Proof of Theorem 3.1

*Proof of Theorem 3.1.* For short hand notation put

$$G_n(k) = \mathbb{P}(S_n \geq k) = \mathbb{P} \left( \bigcup_{j=k}^{n-1} R_n(j) \right).$$

We first consider the case when  $n$  is even. By a simple decomposition

$$\begin{aligned} G_n(k) &= \mathbb{P} \left( \bigcup_{j=k}^{n-1} R_n(j) \right) \\ &= \mathbb{P} \left( \bigcup_{j=k}^{n/2-1} R_n(j) \right) + \mathbb{P} \left( \bigcup_{j=n/2}^{n-1} R_n(j) \setminus \bigcup_{j=k}^{n/2-1} R_n(j) \right) \end{aligned}$$

By Lemma 4.3 the left most term in the last expression is equal to

$$\mathbb{P} \left( \bigcup_{j=k}^{n/2-1} R_{n-2}(j) \right).$$

We can rewrite the last probability as

$$\mathbb{P} \left( \bigcup_{j=k}^{(n-2)-1} R_{n-2}(j) \right) - \mathbb{P} \left( \bigcup_{j=n/2}^{(n-2)-1} R_{n-2}(j) \setminus \bigcup_{j=k}^{n/2-1} R_{n-2}(j) \right).$$

The left most term, by defintion, is equal to  $G_{n-2}(k)$  . Thus, we conclude that

$$\begin{aligned} G_n(k) &= G_{n-2}(k) \\ &\quad + \mathbb{P}(\cup_{j=n/2}^{n-1} R_n(j) \setminus \cup_{j=k}^{n/2-1} R_n(j)) \\ &\quad - \mathbb{P}(\cup_{j=n/2}^{(n-2)-1} R_{n-2}(j) \setminus \cup_{j=k}^{n/2-1} R_{n-2}(j)). \end{aligned}$$

A similar argument shows that

$$\begin{aligned} G_{n-2}(k) &= G_{n-4}(k) \\ &\quad + \mathbb{P}(\cup_{j=n/2-1}^{(n-2)-1} R_{n-2}(j) \setminus \cup_{j=k}^{n/2-2} R_{n-2}(j)) \\ &\quad - \mathbb{P}(\cup_{j=n/2-1}^{(n-4)-1} R_{n-4}(j) \setminus \cup_{j=k}^{n/2-2} R_{n-4}(j)). \end{aligned}$$

Thus

$$\begin{aligned} G_n(k) &= G_{n-4}(k) \\ &\quad + \mathbb{P}(\cup_{j=n/2}^{n-1} R_n(j) \setminus \cup_{j=k}^{n/2-1} R_n(j)) \\ &\quad - \mathbb{P}(\cup_{j=n/2}^{(n-2)-1} R_{n-2}(j) \setminus \cup_{j=k}^{n/2-1} R_{n-2}(j)) \\ &\quad + \mathbb{P}(\cup_{j=n/2-1}^{(n-2)-1} R_{n-2}(j) \setminus \cup_{j=k}^{n/2-2} R_{n-2}(j)) \\ &\quad - \mathbb{P}(\cup_{j=n/2-1}^{(n-4)-1} R_{n-4}(j) \setminus \cup_{j=k}^{n/2-2} R_{n-4}(j)). \end{aligned}$$

Solving the two lines in between we get that they are equal to

$$\mathbb{P}(R_{n-2}(n/2 - 1) \setminus \cup_{j=k}^{n/2-2} R_{n-2}(j)) .$$

A recursive argument up to  $k$  gives

$$\begin{aligned} G_n(k) &= G_{2k}(k) \\ &\quad + \mathbb{P}(\cup_{j=n/2}^{n-1} R_n(j) \setminus \cup_{j=k}^{n/2-1} R_n(j)) \\ &\quad + \sum_{i=k+1}^{n/2-1} \mathbb{P}(R_{2i}(i) \setminus \cup_{j=k}^{i-1} R_{2i}(j)) \\ &\quad - \mathbb{P}(\cup_{j=k+1}^{2k-1} R_{2k}(j) \setminus \cup_{j=k}^k R_{2k}(j)). \end{aligned}$$

Computing the first and last term on the right hand side of the above equality, it gives

$$G_n(k) = R_{2k}(k) + \mathbb{P} \left( \bigcup_{j=n/2}^{n-1} R_n(j) \setminus \bigcup_{j=k}^{n/2-1} R_n(j) \right) + \sum_{i=k+1}^{n/2-1} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=k}^{i-1} R_{2i}(j) \right). \quad (7)$$

This proves *a*) since  $\mathbb{P}(R_{2k}(k)) = m_2^k$ .

Further, since the second term on the right hand side goes to zero as  $n$  diverges, by Lemma 4.4, we conclude that

$$\lim_{n \rightarrow \infty} G_n(k) = \mathbb{P}(R_{2k}(k)) + \sum_{i=k+1}^{\infty} \mathbb{P}(R_{2i}(i) \setminus \bigcup_{j=k}^{i-1} R_{2i}(j)). \quad (8)$$

This proves *c*).

For the probability mass function we have

$$\mathbb{P}(S_n = k) = G_n(k) - G_n(k+1).$$

And solving this equation using (7) we get that  $\mathbb{P}(S_n = k)$  is equal to

$$\begin{aligned} & \mathbb{P}(R_{2k}(k)) - \mathbb{P}(R_{2(k+1)}(k+1)) \\ & - \mathbb{P} \left( \bigcup_{j=n/2}^{n-1} R_n(j) \cap R_n(k) \setminus \bigcup_{j=k+1}^{n/2-1} R_n(j) \right) \\ & - \sum_{i=k+2}^{n/2} \mathbb{P} \left( R_{2i}(i) \cap R_{2i}(k) \setminus \bigcup_{j=k+1}^{i-1} R_{2i}(j) \right) \\ & + \mathbb{P}(R_{2(k+1)}(k+1) \setminus \bigcup_{j=k}^k R_{2(k+1)}(j)). \end{aligned}$$

Computing the right most term in the first line with the last line in the above display, the result is

$$- \mathbb{P}(R_{2(k+1)}(k+1) \cap R_{2(k+1)}(k)).$$

Considering, with some abuse of notation, that the union running over an empty set of indexes is the empty set, we finally get that

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \mathbb{P}(R_{2k}(k)) \\
&- \mathbb{P}\left(\bigcup_{j=n/2}^{n-1} R_n(j) \cap R_n(k) \setminus \bigcup_{j=k+1}^{n/2-1} R_n(j)\right) \\
&- \sum_{i=k+1}^{n/2} \mathbb{P}\left(R_{2i}(i) \cap R_{2i}(k) \setminus \bigcup_{j=k+1}^{i-1} R_{2i}(j)\right).
\end{aligned} \tag{9}$$

This shows *b*).

By Lemma 4.4, term (9) goes to zero as  $n$  diverges. Thus, the limit

$$\lim_{n \rightarrow \infty} \mathbb{P}(S_n = k),$$

exists and is equal to

$$\mathbb{P}(R_{2k}(k)) - \sum_{i=k+1}^{\infty} \mathbb{P}\left(R_{2i}(i) \cap R_{2i}(k) \setminus \bigcup_{j=k+1}^{i-1} R_{2i}(j)\right). \tag{10}$$

This shows *d*).

If  $n$  is odd, the above argument changing  $n/2$  by  $\lfloor n/2 \rfloor$  holds. We conclude that for any positive integer  $n$  we have  $G_{2n+1}(k) = G_{2n}(k)$  and  $\mathbb{P}(S_{2n+1} = k) = \mathbb{P}(S_{2n} = k)$ .

□

## 5.2 Proof of Corollary 3.1.

Note that

$$\mathbb{P}(S_n = 0) = 1 - \mathbb{P}(S_n \geq 1) = 1 - m_2 - a_{1,n},$$

and similarly

$$\lim_{n \rightarrow \infty} \mathbb{P}(S_n = 0) = 1 - m_2 - a_1.$$

But

$$a_1 = \sum_{i=2}^{\infty} \mathbb{P}(R_{2i}(i) \setminus \bigcup_{j=1}^{i-1} R_{2i}(j)).$$

The set in the probability in each term is the set of words  $ww$  where  $w$  is an  $i$ -word without any self-overlap. Namely

$$R_{2i}(i) \setminus \bigcup_{j=1}^{i-1} R_{2i}(j) = \{ww \mid w \in \{S_i = 0\}\} .$$

This establishes the equality in  $b$ ).

Now we prove the last formula. In what follows, the first inequality is just by definition, second one is by Lemma 4.5 and the third one is a simple decomposition.

$$\begin{aligned} G_n(1) &= \mathbb{P} \left( \bigcup_{j=1}^{n-1} R_n(j) \right) \\ &= \mathbb{P} \left( \bigcup_{j=1}^{n/2} R_n(j) \right) \\ &= \mathbb{P} \left( \bigcup_{j=1}^{n/2-1} R_n(j) \right) + \mathbb{P} \left( R_n(n/2) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) \right) . \end{aligned}$$

By Lemma 4.3, the leftmost term in the last display is equal to  $\mathbb{P} \left( \bigcup_{j=1}^{n/2-1} R_{n-2}(j) \right)$ . But applying again Lemma 4.5, this last probability equals to  $\mathbb{P} \left( \bigcup_{j=1}^{(n-2)-1} R_{n-2}(j) \right)$ , which is  $G_{n-2}(1)$ .

It is straightforward to see that

$$\mathbb{P} \left( R_n(n/2) \setminus \bigcup_{j=1}^{n/2-1} R_n(j) \right) = \sum_{w \in \{S_{n/2}=0\}} \mathbb{P}(w)^2 .$$

Thus we conclude that

$$\mathbb{P}(S_n = 0) = \mathbb{P}(S_{n-2} = 0) - \sum_{w \in \{S_{n/2}=0\}} \mathbb{P}(w)^2 .$$

It remains to show the strict inequality in  $b$ ). By the above argument, the probability of the set of  $n$ -strings with some overlap is increasing on  $n$ . Further, the above displays shows that

$$\mathbb{P}(S_n \geq 1) = \mathbb{P}(S_{n-2} \geq 1) + \sum_{w \in \{S_{n/2}=0\}} \mathbb{P}(w)^2 .$$

Now call  $p_1$  and  $p_2$  the two largest  $p_\alpha$ , with  $\alpha \in A$  (allowing multiplicities among the  $p_\alpha$ , tht is  $A$  is considered a multi-set, thus it may happen that  $p_1 = p_2$ ).



That is  $p_1 = \max\{p_\alpha \mid \alpha \in A\}$  and  $p_2 = \max\{p_\alpha \mid \alpha \in A \setminus \alpha_0\}$  where  $p_{\alpha_0} = p_1$ . It follows that, if  $w \in \{S_n = 0\}$  then  $\mathbb{P}(w) \leq p_1^{n-1} p_2$ . Thus, it follows from the last display that

$$\mathbb{P}(S_n \geq 1) \leq \mathbb{P}(S_{n-2} \geq 1) + \mathbb{P}(S_{n/2} = 0) p_1^{n/2-1} p_2 .$$

Since  $\mathbb{P}(S_n = 0)$  is decreasing,

$$\mathbb{P}(S_n \geq 1) \leq \mathbb{P}(S_{n-2} \geq 1) + \mathbb{P}(S_2 = 0) p_1^{n/2-1} p_2 .$$

An iterative argument shows that

$$\mathbb{P}(S_n \geq 1) \leq \mathbb{P}(S_2 = 1) + \mathbb{P}(S_2 = 0) \sum_{j=1}^{n/2-1} p_1^j p_2 .$$

And

$$\lim_{n \rightarrow \infty} \mathbb{P}(S_n \geq 1) \leq \mathbb{P}(S_2 = 1) + \mathbb{P}(S_2 = 0) \frac{p_1 p_2}{1 - p_1} .$$

Since  $p_2 \leq 1 - p_1$  we conclude that

$$\lim_{n \rightarrow \infty} \mathbb{P}(S_n \geq 1) \leq \mathbb{P}(S_2 = 1) + p_1 \mathbb{P}(S_2 = 0) ,$$

observing that  $\mathbb{P}(S_2 = 1) = m_2$ .

### 5.3 Proof of Theorem 3.2.

It follows by Theorem 3.1 that

$$|\mathbb{P}(S_n = k) - \lim_{n \rightarrow \infty} \mathbb{P}(S_n = k)| = |b_{k,n} - b_k| ,$$

which is bounded from above by

$$\max \left\{ \sum_{i=n/2+1}^{\infty} \mathbb{P}(R_{2i}(i) \cap R_{2i}(k)) , \mathbb{P} \left( \bigcup_{j=n/2}^{n-1} R_n(j) \cap R_n(k) \setminus \bigcup_{j=k+1}^{n/2-1} R_n(j) \right) \right\} . \quad (11)$$

Consider firstly the first term in (11). If an  $n$ -string  $w$  belongs to  $R_{2i}(i) \cap R_{2i}(k)$  then it has the form

$$w = w_1 w_2 w_1 w_1 w_2 w_1 ,$$

where  $w_1$  is a  $k$ -string and  $w_2$  is an  $i - 2k$ -string. Therefore

$$\mathbb{P}(w) = \mathbb{P}(w_1)^4 \mathbb{P}(w_2)^2 ,$$

and thus

$$\mathbb{P}(R_{2i}(i) \cap R_{2i}(k)) = \sum_{w_1 \in \mathcal{C}^k} \mathbb{P}(w_1)^4 \sum_{w_2 \in \mathcal{C}^{i-2k}} \mathbb{P}(w_2)^2 = m_4^k m_2^{i-2k} .$$

Summing over  $i$ , we get that the first term in (11) is bounded by

$$\left( \frac{m_4}{m_2^2} \right)^k \frac{m_2^{n/2+1}}{1 - m_2} . \quad (12)$$

Consider now the second term in (11). By duality, the probability in  $b)$  is equivalent to

$$\mathbb{P} \left( \bigcup_{j=1}^{n/2} B_n(j) \cap R_n(k) \setminus \bigcup_{j=n/2+1}^{n-k-1} B_n(j) \right) . \quad (13)$$

Since, by definition,  $B_n(j) \subset B_n(l)$  for all  $l$  multiple of  $j$  one has

$$\bigcup_{j=n/2+1}^{n-k-1} B_n(j) = \bigcup_{j=1}^{n/2-k} B_n(j) \cup \bigcup_{j=n/2+1}^{n-k-1} B_n(j) .$$

Thus, the set in (13) is equal to

$$\bigcup_{j=n/2-k+1}^{n/2} B_n(j) \cap R_n(k) \setminus \bigcup_{j=n/2+1}^{n-k-1} B_n(j) .$$

The above expression implies that it is enough to bound

$$\left( \sum_{j=n/2-k+1}^{n/2-k/2} + \sum_{j=n/2-k/2+1}^{n/2} \right) \mathbb{P}(B_n(j) \cap R_n(k)) = I + II .$$

Consider  $I$ . Since  $2j \leq n - k$  and  $w \in B_n(j)$  then there are at least two complete blocks of length  $j$  at the beginning of  $w$ , and the remaining part of  $w$  has length at least  $k$ . Thus, we can write

$$w = w_b w_b w_l .$$

Further, since  $w \in R_n(k)$ , the first and last block of length  $k$  are equal. Thus

$$w = w_k w_m w_k .$$

The last two descriptions of  $w$  imply that

$$w = w_1 w_2 w_1 w_2 w_3 w_1 .$$

where  $w_1$  has length  $k$  and  $w_2$  has length  $j - k$ . Moreover,  $w_3$  has length  $n - 2j - k$ . Thus, factorizing the measure of  $w$  we have

$$\mathbb{P}(w) = \mathbb{P}(w_1)^3 \mathbb{P}(w_2)^2 \mathbb{P}(w_3) .$$

Recall that  $\rho = \max\{p_\alpha \mid \alpha \in A\}$ . Therefore

$$\mathbb{P}(B_n(j) \cap R_n(k)) \leq \sum_{w_1 \in C^k} \mathbb{P}(w_1)^3 \sum_{w_2 \in C^{j-k}} \mathbb{P}(w_2)^2 \rho^{n-k-2j} = m_3^k m_2^{j-k} \rho^{n-k-2j} .$$

Summing over  $j$  we have

$$\sum_{j=n/2-k+1}^{(n-k)/2} \mathbb{P}(R_n(j) \cap R_n(k)) \leq C_\theta m_2^{n/2} \left( \frac{m_3}{m_2^{3/2}} \right)^k , \quad (14)$$

where  $C_\theta = m_2/(m_2 - \rho^2)$ . Finally, observe that  $m_3/m_2^{3/2} < 1$  is equivalent to  $m_3^{1/3} < m_2^{1/2}$  which is true by Lemma 4.1.

Consider *II*. Take  $w = x_1^n \in B_n(j) \cap R_n(k)$ . Since  $w \in R_n(k)$  one has

$$w = w_k w_m w_k .$$

Since blocks can be read forward or backward, every peace of the string is also periodic (that is, the central peace is in  $B_{n-2k}(j)$ ). So, we can recopilate this and write

$$w = w_1 w_2 w_1 w_2 w_3 w_1 .$$

The length of  $w_1$  is  $k$ . The length of  $w_2$  is  $n - 2k - j$  and the length of  $w_3$  is  $2j + k - n$ . Factorizing the measure of  $w$  we have

$$\mathbb{P}(B_n(j) \cap R_n(k)) \leq \sum_{w_1 \in C^k} \mathbb{P}(w_1)^3 \sum_{w_2 \in C^{n-2k-j}} \mathbb{P}(w_2)^2 \rho^{2j+k-n} \quad (15)$$

$$= m_3^k m_2^{n-2k-j} \rho^{2j+k-n} . \quad (16)$$

Summing over  $j$  we have

$$\sum_{j=(n-k)/2+1}^{n/2} \mathbb{P}(R_n(j) \cap R_n(k)) \leq C'_\theta m_2^{n/2} \left( \frac{m_3}{m_2^{3/2}} \right)^k$$

where  $C'_\theta = (\rho/m_2)^2$ . This ends the proof of  $II$ .

So, as  $C_\theta \geq C'_\theta$ , take  $C = 2C_\theta$ . To end the proof of  $b$ ) we need to show that the right hand side of (12) is less or equal than (14). To this, observe that this is equivalent to show that  $m_4 \leq m_3 m_2^{1/2}$ . But

$$m_4 = \sum_{\alpha \in A} p_\alpha^4 \leq \rho \sum_{\alpha \in A} p_\alpha^3 = \rho m_3 ,$$

and

$$\rho = (\rho^2)^{1/2} \leq (\rho^2 + \sum_{\alpha \in A, p_\alpha \neq \rho} p_\alpha^2)^{1/2} = m_2^{1/2} .$$

This ends the proof of  $a$ ).

The proof of  $b$ ) follows directly from  $a$ ) summing up the error terms in  $a$ ).  $\square$

#### 5.4 Proof of Proposition 3.1.

We first prove  $c$ ). We can write

$$a_k = \left( \sum_{i=k+1}^{2k-1} + \sum_{i=2k}^{\infty} \right) \mathbb{P}(R_{2i}(i) \setminus \cup_{j=k}^{i-1} R_{2i}(j)) = I + II .$$

As in the proof of the first term in (11) with  $n = 4k$  we get

$$II \leq \left( \frac{m_4}{m_2^2} \right)^k \frac{m_2^{2k+1}}{1 - m_2} = m_4^k \frac{m_2}{1 - m_2} .$$

By a direct computation one has

$$I \leq \sum_{i=k+1}^{2k-1} \mathbb{P}(R_{2i}(i)) = \sum_{i=k+1}^{2k-1} m_2^i = \frac{m_2^{k+1} - m_2^{2k}}{1 - m_2} .$$

Thus,  $c$ ) follows since  $m_4^k \leq m_2^{2k}$ .

Proof of  $d$ ).

$$b_k \leq \left( \sum_{i=k+1}^{2k-1} + \sum_{i=2k}^{\infty} \right) \mathbb{P}(R_{2i}(i) \cap R_{2i}(k)) = I + II .$$

As we computed in the proof of an upper bound for (11) when proving Theorem 3.2,  $II$  is

$$m_4^k \sum_{i=0}^{\infty} m_2^i = m_4^k \frac{1}{1 - m_2} .$$

For the leading term in  $I$  we note that

$$R_{2i}(i) \cap R_{2i}(k) = \{ww \mid w \in B_{k+j}(j)\} ,$$

with  $j = i - k$ . Thus, for  $1 \leq j \leq k - 1$  we compute

$$\sum_{w \in B_{k+j}(j)} \mathbb{P}(w)^2 \leq \sum_{w_j \in \mathcal{C}^j} \mathbb{P}(w_j)^{2\lfloor (k+j)/j \rfloor} \rho^{2r} = m_2^j \rho^{2r} .$$

where  $k + j = \lfloor \frac{k+j}{j} \rfloor j + r$  and  $0 \leq r \leq j - 1$ . We conclude that

$$m_2^j \rho^{2r} \leq m_4^j \rho^{2r} \leq m_4^{(k+j-r)/2} \rho^{2r} \leq m_4^{(k+j)/2} .$$

Therefore

$$I \leq m_4^k \frac{m_4 - m_4^k}{1 - m_4} . \quad (17)$$

Proof of  $a)$  and  $b)$ . Similar computations of those done in the proof of  $c)$  and  $d)$  can be done to get an upper bound for the second term in  $a_{k,n}$  and  $b_{k,n}$ .

The second term in  $a_{k,n}$  is bounded by

$$\sum_{i=k+1}^{n/2-1} \mathbb{P}(R_{2i}(i)) = \sum_{i=k+1}^{n/2-1} m_2^i = \frac{m_2^{k+1} - m_2^{n/2}}{1 - m_2} ,$$

and the first one by

$$\sum_{i=n/2}^{n-1} \mathbb{P}(R_n(i)) = \sum_{i=n/2}^{n-1} m_2^i = \frac{m_2^{n/2} - m_2^n}{1 - m_2} .$$

Thus,  $a_{k,n} \leq (m_2^{k+1} - m_2^n)/(1 - m_2)$ .

The first term in  $b_{k,n}$  was bounded in the proof of Theorem 3.2, equation (11) by  $C_\theta m_2^{n/2} \left( \frac{m_3}{m_2^{3/2}} \right)^k$ . The second one is bounded as was done  $b_k$  above.  $\square$

## 5.5 Proof of Proposition 3.2.

$a)$  follows directly from Proposition 3.1  $c)$ .

Now we prove  $b.1)$ . By Corollary 3.1  $b)$ , we have  $a_1 < p_1(1 - m_2) < 1 - m_2 \leq m_2$ . Last inequality follows since  $m_2 \geq 1/2$ .

Now we prove the first sentence and also  $b.2)$ . By definition

$$\begin{aligned}
a_k &= \sum_{i=k+1}^{\infty} \mathbb{P} \left( R_{2i}(i) \setminus \bigcup_{j=k}^{i-1} R_{2i}(j) \right) \\
&= \sum_{i=k+1}^{\infty} \mathbb{P}(R_{2i}(i)) - \sum_{i=k+1}^{\infty} \mathbb{P}(R_{2i}(i) \cap \bigcup_{j=k}^{i-1} R_{2i}(j)) \\
&= \sum_{i=k+1}^{\infty} m_2^i - \sum_{i=k+1}^{\infty} \mathbb{P}(R_{2i}(i) \cap \bigcup_{j=k}^{i-1} R_{2i}(j))
\end{aligned}$$

Bounding the union by the sum we get

$$\begin{aligned}
a_k &\geq \sum_{i=k+1}^{\infty} m_2^i - \sum_{i=k+1}^{\infty} \sum_{j=k}^{i-1} \mathbb{P}(R_{2i}(i) \cap R_{2i}(j)) \\
&= \frac{m_2^{k+1}}{1 - m_2} - \sum_{j=k}^{\infty} \sum_{i=j+1}^{\infty} \mathbb{P}(R_{2i}(i) \cap R_{2i}(j)),
\end{aligned}$$

where the equality was obtained by using Fubini's Theorem. Now, let's take a look at the last term on the previous equation. it can be written as

$$\sum_{j=k}^{\infty} \sum_{i=j+1}^{2j-1} \mathbb{P}(R_{2i}(i) \cap R_{2i}(j)) + \sum_{j=k}^{\infty} \sum_{i=2j}^{\infty} \mathbb{P}(R_{2i}(i) \cap R_{2i}(j)) = I + II$$

Term  $I$  is bounded as in (17):

$$I \leq m_4^k \left( \frac{m_4 - m_4^k}{1 - m_4} \right) \leq \frac{m_4^{k+1}}{1 - m_4}.$$

For the second one we have

$$\begin{aligned}
II &= \sum_{j=k}^{\infty} \sum_{i=2j}^{\infty} \left( \sum_{\omega \in \mathcal{C}^j} \mathbb{P}(\omega)^4 \right) \left( \sum_{\omega \in \mathcal{C}^{i-2j}} \mathbb{P}(\omega)^2 \right) \\
&= \sum_{j=k}^{\infty} \sum_{i=2j}^{\infty} m_4^j m_2^{i-2j} \\
&= \left( \frac{m_4^k}{1 - m_4} \right) \left( \frac{1}{1 - m_2} \right).
\end{aligned}$$

So, we have

$$a_k \geq m_2^k \left( \frac{m_2}{1 - m_2} - \frac{A(k)}{m_2^k} \right),$$

where

$$A(k) = \frac{m_4^k}{1 - m_4} \left( m_4 + \frac{1}{1 - m_2} \right) \geq I + II.$$

Clearly,  $\lim_{k \rightarrow \infty} A(k)/m_2^k = 0$ , and also  $\lim_{k \rightarrow \infty} A(k) = 0$ . Putting on the most left side the lower bound given by Theorem 3.1, we have that

$$\frac{m_2^{k+1}}{1 - m_2} - A(k) \leq a_k \leq \frac{m_2^{k+1}}{1 - m_2},$$

and this proves sharpness. To prove b.2), we just have to notice that since, by hypothesis,  $m_2/(1 - m_2) > 1$ , then there is some  $k_0$  for which:

$$\frac{m_2}{1 - m_2} - \frac{A(k)}{m_2^k} \geq 1, \forall k > k_0.$$

And this concludes the proof.  $\square$

## 6 Non-convergence in probability

In this section we show that  $S_n$  does not converges in probability when  $n$  goes to infinity. Recall that since we are considering non-trivial cases, we have  $\rho < 1$ .

**Proposition 6.1.** *Under the conditions of Theorem 3.1, there is not a random variable  $S$  over  $\mathcal{C}^N$  such that  $S_n$  converge in probability to  $S$ .*

*Proof.* Suppose that  $S_n$  converges to  $S$  in probability. Then, for all  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|S_{n+1} - S_n| < \epsilon) = 1.$$

Consider  $\epsilon < 1$ . Since, by definition  $S_n = n - T_n$  one has

$$\{|S_{n+1} - S_n| < \epsilon\} = \{|T_n - T_{n+1} + 1| < \epsilon\}.$$

Since  $T_n$  is non decreasing and takes only positive integer values

$$\{|T_{n+1} - T_n| < \epsilon\} = \{T_{n+1} = T_n\} = \{S_{n+1} = S_n + 1\}.$$

Conditioning on  $\{T_n = k\}$  we get

$$\mathbb{P}(T_{n+1} = T_n) = \sum_{\alpha \in A} p_\alpha^2 < 1.$$

This ends the proof.  $\square$

**Acknowledgments.** We thank Anatoli Yambarstev, Andrea Vanessa Rocha and Andrei Toom who let us know about Janson's argument. The problem of computing the fluctuations of  $S_n$  was suggested by Pablo Ferrari. The problem of counting the number of non-selfoverlapping strings was suggested by Andrei Toom on the Brazilian School of Probability. M.A. is partially supported by CNPq grant 312904/2009-6. R. L. received CNPq grant 560764/2008-1, between 2008 and 2010. This paper is part of the Projeto Regular Fapesp 2010/19748-7. This work is part of USP project "Mathematics, computation, language and the brain"(Programa da Reitoria da Universidade de São Paulo de Incentivo à Pesquisa - Projeto MaCLinC - Matemática, Computação, Linguagem e Cérebro, Processo USP no. 2011.1.9367.1.5).

## References

- [1] M. Abadi, *Exponential approximation for hitting times in mixing processes*, Math. Phys. Elec. J. (7) 2 (2001).
- [2] M. Abadi, *Sharp error terms and necessary conditions for exponential hitting times in mixing processes*, Ann. Probab. (32) 1A (2004), 243–264.
- [3] M. Abadi, *Hitting, returning and the short correlation function*, Bull. Braz. Math. Soc. (37) 4 (2006), 593-609.
- [4] M. Abadi and L. Cardeno, *Renyi entropies and large deviations for the overlapping function*, Preprint.
- [5] M. Abadi and B. Saussol (2010) *Hitting and returning into rare events for all alpha-mixing processes* [http://arxiv.org/PS\\_cache/arxiv/pdf/1003/1003.4856v1.pdf](http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.4856v1.pdf)
- [6] M. Abadi and S. Vaienti, *Large Deviations for Short Recurrence* Disc. Cont. Dyn. Syst. 21 (2008), no. 3, 729–747
- [7] M. Abadi and N. Vergne, (2008) *Sharp error terms for return time statistics under mixing condition*, 21 (2008), no. 3, 729–747
- [8] V. Afraimovich, J.-R. Chazottes and B. Saussol, *Point-wise dimensions for Poincar recurrence associated with maps and special flows*, Discrete Contin. Dyn. Syst. 9 (2003), no. 2, 263–280. (2003).
- [9] P. Collet, A. Galves and B. Schmitt, *Repetition times for gibbsian sources*, Nonlinearity 12 (1999), 1225–1237.
- [10] A. Galves and B. Schmitt, *Inequalities for hitting times in mixing dynamical systems*, Random Comput. Dyn. 5 (1997), 337–348.
- [11] N. Haydn and S. Vaienti, *The Renyi entropy function and the large deviation of short return times*, Preprint. To appear in Ergodic Theory and Dynamical Systems.



- [12] N. Haydn and S. Vaienti, *The limiting distribution and error terms for return time of dynamical systems*, Discrete and Continuous Dynamical Systems 10 (2004), 584–616.
- [13] M. Hirata, *Poisson law for Axiom A diffeomorphisms*. Ergodic Theory Dynam. Systems 13, no. 3, (1993) 533–556.
- [14] M. Hirata, B. Saussol and S. Vaienti, *Statistics of return times: a general framework and new applications*, Comm. Math. Phys. 206 (1999), 33–55.
- [15] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.* **53**, (1947) 1002-1010.
- [16] A. Katok and B. Hasselblatt, "Introduction to the modern theory of dynamical systems," Encyclopedia of Math. and its Applications, 54, Cambridge Univ. Press 1995.
- [17] G. Reinert and S. Schbath, *Compound Poisson approximations for occurrences of multiple words*. Statistics in Genetics and Molecular Biology, (F. Seillier, ed.). IMS Lecture Notes-Monograph Series. Vol. 33, 1999.
- [18] G. Reinert and S. Schbath, . *Compound Poisson and Poisson process approximations for occurrences of multiple words in markov chains*. J. Comp. Biol. 5 (1998) 223-254.
- [19] A. Rocha. *Substitution Operators*. Phd Thesis, Universidade Federal de Pernambuco (2009).
- [20] E. Roquain, and S. Schbath, *Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain*. Adv. Appl. Prob. 39 (2007) 1-13.
- [21] B. Saussol, S. Troubetzkoy and S. Vaienti, *Recurrence, dimensions and Lyapunov exponents*, J. Stat. Phys. 106 (2002), 623–634.
- [22] K. Zyczkowski, *Renyi extrapolation of Shannon entropy* Open Syst. Inf. Dyn. 10, (2003) 297-310; with 2005 corrigendum: <http://www.cft.edu.pl/~karol/pdf/Zy03b.pdf>